HSR Health Services Research

DOI: 10.1111/j.1475-6773.2011.01305.x
RESEARCH BRIEF

Accuracy of Data Entry of Patient Race/Ethnicity/Ancestry and Preferred Spoken Language in an Ambulatory Care Setting

Kristen M.J. Azar, Maria R. Moreno, Eric C. Wong, Jessica J. Shin , Christy Soto , and Latha P. Palaniappan

Objective. To describe data collection methods and to audit staff data entry of patient self-reported race/ethnicity/ancestry and preferred spoken language (R/E/A/L) information.

Data Source/Study Setting. Large mixed payer outpatient health care organization in Northern California, June 2009.

Study Design. Secondary analysis of an audit planned and executed by the Department of Clinical Services.

Data Collection/Extraction Methods. We analyzed concordance between patient written responses and staff data entry.

Principal Findings. The data entry accuracy rate across questions was high, ranging from 92 to 97 percent. Inaccuracies were due to human error (62 percent), flaws in system design (2 percent), or some combination of both (35 percent).

Conclusions. This study highlights the high accuracy of patient self-reported R/E/A/L data entry and identifies some areas for improvement in staff training and technical system design to facilitate further progress.

Key Words. Racial/ethnic differences in health and health care, health care organizations and systems, demography, survey research and questionnaire design, quality of care/patient safety (measurement)

BACKGROUND

Accurate collection of race/ethnicity/ancestry and preferred spoken language (R/E/A/L) patient information is a fundamental building block for disparities research and quality improvement efforts in a health care setting. Patient R/E/A/L data are increasingly being used to evaluate population

outcomes, measure health care disparities, and improve quality of care. Accurate R/E/A/L data collection allows health care organizations to better understand population health and outcomes. In addition, patient race/ethnicity and language reporting is mandatory to state agencies, such as the California Office of Statewide Health Planning and Development (OSHPD), and federal agencies, such as Medicare (Medicare Improvements for Patients and Providers Act of 2008). Most recently, the Patient Protection and Affordable Care Act (PPACA) (PPACA and Education Reconciliation Act 2010, March 23, 2010) calls for reliable and enhanced collection and reporting of patient race/ethnicity and language data to ensure accurate information on the health status and health care needs of all Americans.

Advances in health information technology such as the implementation of electronic health records (EHRs) provide a promising approach to collecting and utilizing patient R/E/A/L information. Even as health care organizations shift from paper to EHRs, demographic data are often collected using a paper format and then entered by staff into the EHR (or Practice Management System). The modern health care environment involves frequent interaction between human (people, tasks, and organization) and system (technologies, equipment, and physical settings of work) aspects that impact performance outcomes (Harrison, Henriksen, and Hughes 2007) such as data entry accuracy. Regular audits of new data entry processes may be helpful to assess and improve data collection efforts (Peabody et al. 2004), by identifying opportunities for improvement in both the human and system aspects of accurate data entry.

Although there is extensive literature on the accuracy of data from disease registries and clinical trial databases (McKee 1993; Wagner, and Hogan 1996; Hogan, and Wagner 1997; Brennan, and Stead 2000; Arts, De Keizer, and Scheffer 2002; Warsi, White, and McCulloch 2002; Peabody et al. 2004; Hobson, Khemani, and Singh 2005), there have been surprisingly few studies on measuring staff data entry accuracy of patient self-reported R/E/A/L in patient registration databases. This article seeks to describe an audit of staff data entry of patient R/E/A/L data in a large outpatient clinical setting, to

---

Address correspondence to Latha Palaniappan, M.D., M.S., Palo Alto Medical Foundation Research Institute (PAMFRI), 795 El Camino Real, Ames Building, Palo Alto, CA 94301; e-mail: lathap@pamf.org. Kristen M.J. Azar, R.N., M.S.N./M.P.H., Eric C. Wong, M.S., Jessica J. Shin, B.A., and Latha P. Palaniappan, M.D., M.S., are with the Palo Alto Medical Foundation Research Institute (PAMFRI), 795 El Camino Real, Ames Building, Palo Alto, CA, 94301. Maria R. Moreno, M.P.H. and Christy Soto, B.S., are with the Sutter Health Institute for Research and Education (SHIRE), 345 California Street, Suite 2000, San Francisco, CA.

understand the root causes of data entry errors, and to make recommendations for other organizations embarking on the process of patient demographic data collection.

CASE DESCRIPTION

The Palo Alto Medical Foundation (PAMF) has been using EHRs since 2000, and it began collecting patient self-reported R/E/A/L information in May 2008 using a paper questionnaire closely modeled after the relevant questions on the U.S. Census 2000 (and 2010). The questionnaire, described briefly here and in detail elsewhere (Palaniappan et al. 2009), consists of questions pertaining to race, Hispanic origin, ancestry, preferred spoken language, and need for interpreter services. The questionnaire is distributed and collected from all patients by the front desk staff at patient check-in (Palaniappan et al. 2009). For patients with limited English proficiency (LEP) (<7 percent of the clinic population), interpreter services are available to assist patients. Interpreter services are openly advertised (in print) at the registration desks in 20 major languages. Patients complete the R/E/A/L questionnaire (available at http://www.pamf.org/real/) privately in the waiting area and return it to the front desk staff for data entry. Most patients are willing to provide the requested information (> 90 percent). A few patients choose to leave race information blank altogether (<7 percent) or choose the "I prefer not to answer" option (<3 percent).

All clinic administrative staff received training prior to the implementation of the new R/E/A/L data collection procedures via a "train the trainer" approach. Department managers participated in a mandatory, standardized, 4-hour seminar on the rationale and protocol for R/E/A/L data collection administered by trainers from the Sutter Health Institute for Research and Education (SHIRE). Clinic administrative staff were trained to enter patient R/E/A/L responses into the patient registration software exactly as the responses appear on the questionnaire. The questionnaire consisted of both checkbox and free response questions. The presentation on the user interface on computer registration screen was only slightly different than the paper questionnaire. All of the data fields were represented in the same order. Check boxes and lists (e.g. Race, Hispanic Origin, and Interpreter Services) were replicated on the computer screen. Free response questions on the questionnaire (e.g. Ancestry and Preferred Spoken Language) correspond to free text fields in the registration screen that are

linked to extensive drop-down menus with text auto-complete functionality to aid staff in entering the information quickly and efficiently. Slight variations between the paper questionnaire and the registration screen are described in detail in the sections to follow.

METHODS

Setting

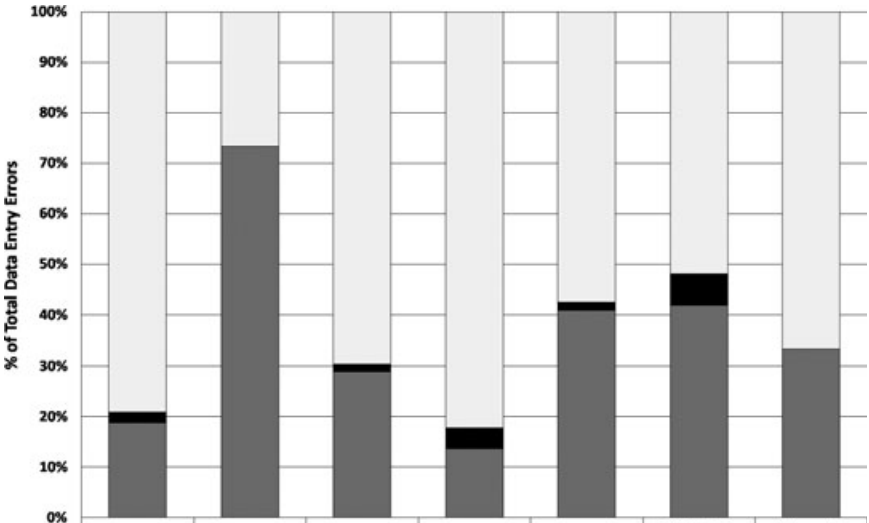
Palo Alto Medical Foundation, a Sutter Health affiliate, is a large multispecialty ambulatory care organization with health care clinics throughout Northern California and the San Francisco bay area. PAMF delivers health care coverage to approximately 15 percent of the general population in four Northern California counties (Alameda, San Mateo, Santa Clara, Santa Cruz), with 35 medical clinics and over 830 clinic administrative staff. Across PAMF, there are over 650,000 active patients with approximately 2.3 million patient visits per year, characterized by wide racial/ethnic and linguistic diversity. To date, of the active PAMF patients who have self-reported their R/E/A/L (65 percent of all active patients), 54 percent self-identify as White/Caucasian, 30 percent self-identify as one of the six major Asian racial/ethnic groups (12 percent Asian Indian, 11 percent Chinese, 3 percent Filipino, 1 percent each of Japanese, Korean, and Vietnamese), 2 percent identify themselves as Black/African American, and approximately 10 percent self-identify as Hispanic/Latino.

Design

The PAMF clinical services team audited staff data entry of patient self-reported R/E/A/L data for 1 week: June 15–June 19, 2009. A complete week was selected to moderate potential bias resulting from the day of the week. The organization's purpose for this audit was quality improvement, to assess the accuracy of data entry, and to inform interventions to improve data entry accuracy. The audit was designed for quality improvement (not research) by the Department of Clinical Services, and, therefore, conclusions about the results should be interpreted with caution. We have secondarily analyzed the results of this audit to provide generalizable lessons regarding errors resulting from human and system interactions in R/E/A/L data entry for other organizations attempting a similar approach. Two auditors (C.S. and J.S.) manually reviewed patient paper

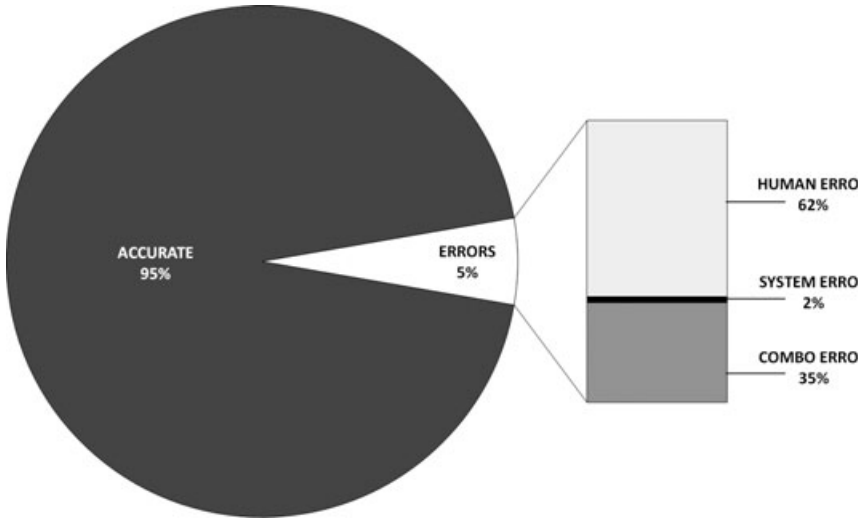
questionnaire responses (completed in the waiting room prior to the physician visit) and compared them with the values entered by staff into the patient's electronic registration record. The questionnaire was composed of five questions on race, Hispanic origin, ancestry, preferred spoken language, and interpreter services. Patients can respond with up to two races and two ancestries; yes or no Hispanic origin and interpreter services; and one preferred spoken language. Questions from the paper questionnaire are linked to seven audited database fields: Race1, Race2, Hispanic Origin, Ancestry1, Ancestry2, Preferred Spoken Language, and need for Interpreter Services (Palaniappan et al. 2009). If differences between questionnaire response and electronic entries were discovered, the auditors made appropriate corrections directly in the electronic patient registration system. These corrections were automatically monitored and classified (Figure 1). Accuracy rates were calculated for each

Figure 1: Distribution of Data Entry Errors by Contributing Factor across R/E/A/L Database Fields. HUMAN, error resulting from human behaviors; SYSTEM, error resulting from suboptimal system features; COMBO, error resulting from combined contribution



	Race 1	Race 2	Hispanic Origin	Ancestry 1	Ancestry 2	Preferred Language	Interpreter Services
HUMAN	79%	27%	70%	82%	58%	52%	67%
SYSTEM	2%	0%	2%	4%	2%	6%	0%
COMBO	19%	73%	29%	14%	41%	42%	33%

Figure 2: Data Entry Errors by Contributing Factor. Due to rounding, percentages sum to less than 100 percent; HUMAN: error resulting from human behaviors; SYSTEM: error resulting from suboptimal system features; COMBO: error resulting from combined contribution



question. Errors were classified as human errors, system errors, or some combination of both (see Figure 2).

FINDINGS

Overall Observations

The total number of eligible questionnaires for analysis was 1,451, after 310 questionnaires were excluded as ineligible by the Department of Clinical Services, mostly due to multiple patient visits within the same week. These questionnaires were excluded due to anticipated systematic differences in these frequent use patients. The accuracy rate for questionnaires (i.e., all seven possible responses on the questionnaire were accurately entered into the electronic registration system) was 81 percent. Among the questionnaires containing any type of error (19 percent of all questionnaires), most had only one error (59 percent), 22 percent had two errors, and 19 percent had three or more errors. The accuracy rate across questions was uniformly high, ranging

from 92 percent (Ancestry 2) to 97 percent (Race1). The mean accuracy rate across all questions was 95 percent (see Figure 2). The types of errors varied across questions and were categorized as resulting from human-related error (62 percent), a flaw purely in system design (2 percent), or some combination of both (35 percent) (see Figure 2).

Observations of Human Contribution to Data Entry Errors

Human errors (62 percent) account for the majority of errors made in the Race 1, Hispanic Origin, and Ancestry fields (see Figure 1). Human errors in R/E/A/L data entry most often occurred when the clinic administrative staff (1) entered a completely different value in the patient's electronic registration record than what the patient indicated on the questionnaire (38 percent) or (2) did not enter the patient's response at all (62 percent). Differences between patient written responses and electronic entries were the most frequently occurring errors in the Race1 (54 percent), Ancestry 1 (66 percent), and Ancestry 2 (44 percent). Situations where the written patient response was not entered at all into the electronic registration record constitute a large portion of total errors in the Hispanic Origin (52 percent) and Interpreter Services (60 percent) fields.

Observations of System Contribution to Data Entry Errors

Other data entry errors (2 percent) stemmed from flaws in the electronic system set-up and structure of the paper questionnaire, which they were revealed in audit analysis. For example, administrative staff did not have editing privileges in the electronic system for these R/E/A/L values, and they were not able to update or correct database fields that had pre-existing data (with the exception of Preferred Spoken Language and Interpreter Services). Lack of editing privileges accounted for the majority of all system errors.

Observations of Combined Contribution to Errors

Combination (combo) errors accounted for 35 percent of all errors. Most often, small inconsistencies between the paper questionnaire and the electronic system led to confusion in data entry by the staff. For example, when a patient returns a blank paper questionnaire, the clinic administrative staff is instructed to enter the term "Left Blank" in the electronic system according to the protocol. Although the Race, Hispanic Origin, and Ancestry questions

had an option for "Left Blank" in the electronic system, the Preferred Language and Interpreter Services database fields did not have a "Left Blank" option in the drop-down menu of choices. These inconsistencies in the system account for the vast majority (90 percent) of all combo errors and led to human error. In some instances, administrative staff did not utilize the "Left Blank" option when it was available, resulting in 58 percent of all combo errors made. In the Interpreter Services database field, where "Left Blank" was not an option, errors in which the staff enters a response when none was indicated by the patient account for 15 percent of all combo errors and 33 percent of total errors in that particular field. In addition, the Preferred Language field is programmed as a mandatory field in the electronic system. Therefore, when a patient does not write a response for this question on the paper questionnaire, the administrative staff is more likely to enter a possible patient response based on assumption, consequently resulting in 42 percent of all of the errors for the Preferred Spoken Language field.

DISCUSSION

Data entry audits are important to ensure valid and reliable data, which enable continuous quality improvement efforts. This audit revealed that the overall accuracy rate of data entry for R/E/A/L is high, at 92–97 percent for each question. The mean accuracy rate across all questions was 95 percent. Existing literature on accuracy of race/ethnicity and language data mainly compares administrative data with self-report (Boehmer et al. 2002; Kressin et al. 2003) or externally completed survey data (Arday et al. 2000) and report much lower rates of accuracy. Our study has taken previous work in R/E/A/L data collection a step further by examining data entry accuracy within an ambulatory care system that already collects *self-reported* R/E/A/L information (Palaniappan et al. 2009). This article is the first to our knowledge that examines an audit process for the collection of self-reported R/E/A/L in an ambulatory care setting.

Accurate R/E/A/L data entry allows an organization to better understand its patients and provide targeted services and prevention efforts to more effectively address the needs of the community it serves. For example, Asian Indians make up a substantial portion (12 percent) of the PAMF patient population and are at increased risk for cardiovascular disease due to certain genetic, cultural, and environmental risk factors (Palaniappan, Wang, and Fortmann 2004). The collection of granular and accurate R/E/A/L has lead

to the creation of a culturally sensitive South Asian consult service that specifically provides preventive cardiology services to South Asian PAMF patients.

Although the data entry accuracy of Preferred Spoken Language and Interpreter Services in this instance was found to be high (94 and 95 percent, respectively), the consequences of inaccurate language and interpreter data entry are especially detrimental to the provision of high-quality patient care. Inability to capture accurate language data may mislead the clinic to assume that there are more English speakers than there really are, resulting in insufficient allocation of language or interpreter services. This is especially burdensome when the clinic needs to determine language needs for health care delivery, communications, and written informed consent. From the patient perspective, pervasive language barriers can easily discourage patients from seeking timely medical care. Not surprisingly, patients with LEP are reluctant to seek services from providers who are unable to communicate effectively with them (Mateo, Gallardo, and Huang 2009; Hunt, and de Voogd 2007). Currently, 10 percent, or approximately 66,000 of all PAMF patients who have completed the survey, report a language other than English as their preferred spoken language. The implications of inaccurate data entry in Interpreter Services in a largely diverse patient community could potentially impact the quality of care.

With health information technology infiltrating all aspects of the patient encounter, the technological interface between humans and systems, in this case regarding patient R/E/A/L data entry, has become paramount. Common errors in data entry in this study were associated with human behavior (62 percent), flaws in the system (2 percent), or a combination of the two (35 percent). Although the audit process uncovered that there was considerable human error, flaws in the technical system design also resulted in errors. By clearly identifying these sources of error, appropriate interventions can be implemented in the form of targeted training and system changes to increase data accuracy.

Proposed training enhancements and system changes are shown in Table 1. Retraining in the importance of accurate data entry, including the database field option for "Left Blank", would improve overall data accuracy by more than 3 percent. As described above, clinic administrative staff were trained via a "train the trainer" model. The extent to which front line staff was subsequently trained by their department managers is unclear. A better model, which we are currently implementing, may be direct training of all staff via a mandatory online training module. The revised training will

Table 1: Contributing Factors and Corresponding Intervention(s)

Contributing Factor	*Intervention*	*Examples*
HUMAN	Targeted, staff training through a mandatory online training module	Directly train clinic administrative staff by requiring them to complete an online R/E/A/L data collection training module through the institution's preexisting staff training infrastructure
		Educate staff on the importance of exact input of patient responses in the electronic system
		Educate staff that no database fields should be left empty in the electronic system
		Educate staff on the availability of "Left Blank" as a valid and necessary option in the electronic system
SYSTEM	System modifications to registration software and paper questionnaire	Give clinic administrative staff system permissions to edit R/E/A/L database fields
		Ensure consistency of the electronic field options that includes the addition of the option for "Left Blank" in both the Preferred Language and Interpreter Services electronic fields
		Ensure consistency between paper questionnaire response options and electronic field response options
		Change the system to prevent entries in Race 2 or Ancestry 2 in electronic system unless Race 1 or Ancestry 1 are completed when more than one response is indicated on the paper questionnaire

HUMAN, error resulting from human behaviors; SYSTEM, error resulting from suboptimal system features.

highlight common data entry errors. System changes include greater consistency across questions, and the ability to edit data fields as appropriate.

Given the majority of data entry errors (62 percent) were due to human error, the ideal system might involve patients entering their R/E/A/L information directly. This can be done with a kiosk or online patient portal in which the patient can directly interface with the electronic system and enter their R/E/A/L information. This would eliminate the human-related error that exists when administrative staff serves as the intermediary. Although the majority of R/E/A/L data is still collected during a face-to-face patient encounter, health plans such as Aetna, HealthPartners, and UnitedHealth Group, as well as large institutional health care providers, are increasingly utilizing web-based patient portals to allow patients the opportunity to self-enter R/E/A/L information (National Health Plan 2008).

CONCLUSION

Overall, this study highlights the high data entry accuracy of patient self-reported R/E/A/L information from paper questionnaires and identifies some areas for improvement in staff training and technical systems. The overall goal for staff data entry accuracy at PMF is 99 percent for each question. We have identified several areas of social and technical improvement, which are currently being implemented. For more detailed information on training protocols and R/E/A/L data collection, please visit http://www.pamf.org/real/.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: All authors comply with the editorial policies and do not report any competing interests. This manuscript is an original work of authorship that is not under simultaneous consideration elsewhere, and the final version has been read and approved by all individuals named as authors. This work presents novel information that differs substantially from that presented in works published by the authors previously.

In addition, each author has (1) contributed significantly to the work's conception, design, and analysis; (2) participated in the writing or critical revision of the article in a manner sufficient to establish ownership of the intellectual content; and (3) read and approved the version of the manuscript being submitted.

The authors acknowledge the assistance and material support of the Department of Clinical Services, Palo Alto Medical Foundation (Palo Alto Division) and specifically thank Mr. Michael Fagan and Ms. Genevieve Buchanan for their support and administration of this project.

Disclosures: None.

Disclaimers: None.

REFERENCES

Arday, S., D. Arday, S. Monroe, and J. Zhang. 2000. "HCFA's Racial and Ethnic Data: Current Accuracy and Recent Improvements." *Health Care Financing Review* 21 (4): 107.

Arts, D. G., N. F. De Keizer, and G. J. Scheffer. 2002. "Defining and Improving Data Quality in Medical Registries: A Literature Review, Case Study, and Generic Framework." *Journal of the American Medical Informatics Association* 9 (6): 600–11.

Boehmer, U., N. R. Kressin, D. R. Berlowitz, C. L. Christiansen, L. E. Kazis, and J. A. Jones. 2002. "Self-Reported vs Administrative Race/Ethnicity Data and Study Results." *American Journal of Public Health* 92 (9): 1471–2.

Brennan, P. F., and W. W. Stead. 2000. "Assessing Data Quality: From Concordance, Through Correctness and Completeness, to Valid Manipulatable Representations." *Journal of the American Medical Informatics Association* 7 (1): 106–7.

Harrison, M. I., K. Henriksen, and R. G. Hughes. 2007. "Improving the Health Care Work Environment: A Sociotechnical Systems Approach." *Joint Commission Journal on Quality and Patient Safety* 33 (11 Suppl): 1, 3–6.

Hobson, J. C., S. Khemani, and A. Singh. 2005. "Prospective Audit of the Quality of ENT Emergency Clinic Notes before and after Introduction of a Computerized Template." *Journal of Laryngology and Otology* 119 (4): 264–6.

Hogan, W. R., and M. M. Wagner. 1997. "Accuracy of Data in Computer-Based Patient Records." *Journal of the American Medical Informatics Association* 4 (5): 342–55.

Hunt, L. M., and K. B. de Voogd. 2007. "Are Good Intentions Good Enough? Informed Consent without Trained Interpreters." *Journal of General Internal Medicine* 22 (5): 598–605.

Kressin, N. R., B. H. Chang, A. Hendricks, and L. E. Kazis. 2003. "Agreement between Administrative Data and Patients Self-Reports of Race/Ethnicity." *American Journal of Public Health* 93 (10): 1734–9.

Mateo, J., E. V. Gallardo, and V. Y. Huang. 2009. "*Providing Health Care to Limited English Proficiency (LEP) Patients: A Manual of Promising Practices. California Primary Care Association.*" Sacramento, CA: Centers for Medicare and Medicaid Services/AKA Health Care Financing Administration[accessed on March 23, 2010]. Available at: http://www.hhs.gov/ocr/civilrights/resources/specialtopics/lep/providinghealthcaretoleppdf.pdf.

McKee, M. 1993. "Routine Data: A Resource for Clinical Audit?" *Quality in Health Care* 2 (2): 104–11.

H.R. 6331--110th Congress: Medicare Improvements for Patients and Providers Act of 2008. July 15, 2008. Public Law 110-275 § 118, 110th Cong., 2d sess.

National Health Plan Collaborative. 2008. "*Toolkit to Reduce Racial & Ethnic Disparities in Health Care.*" National Health Plan Collaborative [accessed March 23, 2010]. Available at: http://www.rwjf.org/files/research/nhpctoolkit.pdf

Palaniappan, L., Y. Wang, and S. P. Fortmann. 2004. "Coronary Heart Disease Mortality for Six Ethnic Groups in California, 1990–2000." *Annals of Epidemiology* 14 (7): 499–506.

Palaniappan, L. P., E. C. Wong, J. J. Shin, M. R. Moreno, and R. Otero-Sabogal. 2009. "Collecting Patient Race/Ethnicity and Primary Language Data in Ambulatory Care Settings: A Case Study in Methodology." *Health Services Research* 44 (5 Pt 1): 1750–61.

Patient Protection and Affordable Care Act (PPACA) and Education Reconciliation Act. March 23, 2010 [accessed on March 23, 2010]. Available at: http://s3.amazonaws.com/thf_media/2010/pdf/ppaca-consolidated.pdf.

Peabody, J. W., J. Luck, S. Jain, D. Bertenthal, and P. Glassman. 2004. "Assessing the Accuracy of Administrative Data in Health Information Systems." *Medical Care* 42 (11): 1066–72.

Wagner, M. M., and W. R. Hogan. 1996. "The Accuracy of Medication Data in an Outpatient Electronic Medical Record." *Journal of the American Medical Informatics Association* 3 (3): 234–44.

Warsi, A. A., S. White, and P. McCulloch. 2002. "Completeness of Data Entry in Three Cancer Surgery Databases." *European Journal of Surgical Oncology* 28 (8): 850–6.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.